Jan Vlček; Ladislav Lukšan

Variationally-derived limited-memory methods for unconstrained optimization

In: Jan Chleboun and Petr Přikryl and Karel Segeth (eds.): Programs and Algorithms of Numerical Mathematics, Proceedings of Seminar. Dolní Maxov, June 6-11, 2004. Institute of Mathematics AS CR, Prague, 2004. pp. 273–278.

Persistent URL: http://dml.cz/dmlcz/702804

VARIATIONALLY-DERIVED LIMITED-MEMORY METHODS FOR UNCONSTRAINED OPTIMIZATION*

Jan Vlček, Ladislav Lukšan

1. Introduction

Variable metric (VM) methods, see [2], [5], for unconstrained minimization, are the most popular iterative methods for small and medium-size problems, since they are simple and numerically efficient and have good convergence properties. Starting with an initial point $x_1 \in \mathcal{R}^N$, they generate a sequence $x_k \in \mathcal{R}^N$, $k \geq 1$, by the process $x_{k+1} = x_k + t_k d_k$, where $d_k \in \mathcal{R}^N$ is a direction vector and $t_k \geq 0$ is a stepsize.

We assume that the problem function $f : \mathcal{R}^N \to \mathcal{R}$ is uniformly convex and has bounded second-order derivatives and denote $f_k = f(x_k)$, $g_k = \nabla f(x_k)$, $s_k = x_{k+1} - x_k$ and $y_k = g_{k+1} - g_k$, $k \geq 1$. VM methods use symmetric positive definite matrices H_k, $k \geq 1$; usually $H_1 = I$ and H_{k+1} is obtained from H_k by a rank-two VM update to satisfy the quasi-Newton condition $H_{k+1} y_k = s_k$.

We will investigate the line search methods satisfying

$$d_k = -H_k g_k, \quad f_{k+1} - f_k \leq \varepsilon_1 t_k g_k^T d_k, \quad g_{k+1}^T d_k \geq \varepsilon_2 g_k^T d_k, \tag{1}$$

$k \geq 1$, where $0 < \varepsilon_1 < 1/2$, $\varepsilon_1 < \varepsilon_2 < 1$. The most efficient VM methods belong to the Broyden class with parameter η (we often omit index k and replace $k+1$ by symbol $+$) $H_+ = H + ss^T/b - Hyy^T H/a + (\eta/a)\,[(a/b)s - Hy]\,[(a/b)s - Hy]^T$, $a = y^T H y$, $b = s^T y$. For $\eta = 0$ we obtain the DFP update, for $\eta = 1$ the BFGS update.

2. Shifted variable metric methods

Here matrices H_k have the form $H_k = \zeta_k I + A_k$, $k \geq 1$, see [9], [10], where $\zeta_k > 0$ and A_k are symmetric positive semidefinite; usually $A_1 = 0$ and A_{k+1} is obtained from A_k by a VM update to satisfy shifted analogy of the quasi-Newton condition

$$A_+ y = \varrho \tilde{s}, \qquad \tilde{s} = s - \zeta_+ y. \tag{2}$$

If $\varrho = 1$, matrix H_+ satisfies the quasi-Newton condition $H_+ y = s$. We use non-unit values of ϱ only for variationally-derived limited-memory methods. We denote

$$a = y^T H y, \;\; \bar{a} = y^T A y, \;\; \hat{a} = y^T y, \;\; b = s^T y, \;\; \bar{b} = y^T A B s, \;\; \tilde{b} = \tilde{s}^T y, \;\; B = H^{-1}.$$

*This work was supported by the Grant Agency of the Czech Academy of Sciences, project code IAA1030405, and with the subvention from Ministry of Education of the Czech Republic, project code MSM 242200002.

As in standard VM methods, we obtain the shifted analogy of the Broyden class for $\tilde{b} > 0$ (which implies $\tilde{s} \neq 0$, $y \neq 0$; if $\bar{a} = 0$, we simply omit the last two terms)

$$A_+ = A + \varrho \frac{\tilde{s}\tilde{s}^T}{\tilde{b}} - \frac{Ayy^TA}{\bar{a}} + \frac{\eta}{\bar{a}} \left(\frac{\bar{a}}{\tilde{b}} \tilde{s} - Ay \right) \left(\frac{\bar{a}}{\tilde{b}} \tilde{s} - Ay \right)^T . \tag{3}$$

For $\eta = 0$ we obtain the shifted DFP update, for $\eta = 1$ the shifted BFGS update.

Theorem 2.1. *Let A be positive semidefinite, $\eta \geq 0$ and $\mu \triangleq \zeta_+ \hat{a}/b < 1$. Then matrix A_+ given by (3) is positive semidefinite.*

We will suppose that $\eta \geq 0$ and $\mu \in (0, 1)$. In the simplest strategy of the shift parameter μ determination, its value remains the same in all iterations (e.g. $0.20 \leq \mu \leq 0.25$). If $\mu \geq 1/2$, then the convergence is usually lost (the shifted DFP method is an exception). Analysis of the shifted BFGS method leads to formula

$$\mu = \sqrt{\zeta \hat{a}/a} \Big/ \left(1 + \sqrt{1 - b^2/(\hat{a}|s|^2)} \right) , \tag{4}$$

which gives good results, with the exception of the first five to ten iterations, when it should be corrected, e.g. $\mu = \min(\max(\sqrt{1 - \bar{a}/a}/(1 + \sqrt{1 - b^2/(\hat{a}|s|^2)}\,), 0.2), 0.8)$.

2.1. Global convergence

Assumption 2.1. *The objective function $f : \mathcal{R}^N \to \mathcal{R}$ is uniformly convex and has bounded second-order derivatives (i.e. $0 < \underline{G} \leq \underline{\lambda}(G(x)) \leq \overline{\lambda}(G(x)) \leq \overline{G} < \infty$, $x \in \mathcal{R}^N$, where $\underline{\lambda}(G(x))$ and $\overline{\lambda}(G(x))$ are the lowest and the greatest eigenvalues of the Hessian matrix $G(x)$).*

Assumption 2.2. *Parameters ϱ_k and μ_k of the shifted VM method are uniformly positive and bounded, in the sense that $0 < \underline{\varrho} \leq \varrho_k \leq \overline{\varrho}$, $0 < \underline{\mu} \leq \mu_k \leq \overline{\mu} < 1$, $k \geq 1$.*

Theorem 2.2. *Consider the shifted variable metric method (3) satisfying Assumption 2.2 with $\underline{\mu}$ sufficiently small and suppose that the line search method fulfils (1). Let the objective function satisfy Assumption 2.1. If $\eta \in [0, 1]$ and $\mu^2 \leq \zeta \hat{a}/a$ or $\mu = \underline{\mu}$ (e.g. if $\underline{\mu}^2 > \zeta \hat{a}/a$), then $\liminf_{k\to\infty} |g_k| = 0$ (global convergence).*

2.2. Computational experiments

We use a collection of 92 relatively difficult problems ([7], Test 28) and the final precision $\|g(x^\star)\|_\infty \leq 10^{-6}$. In the tables, NIT is the total number of iterations, NFV the number of function (or gradient) evaluations and 'Fail' denotes the number of problems where NFV reached its limit. Table 1 demonstrates an influence of the constant parameter μ on the shifted BFGS method. In Table 2 we use choice (4) of μ with the mentioned corrections in the first six iterations. The first three rows contain results for the shifted BFGS method (SBFGS, $\eta = 1$), the method (3) with $\eta = 2$ (SBC2) and the shifted DFP method (SDFP, $\eta = 0$).The last four rows contain

μ	0.22	0.32	0.42	0.48	0.50	0.52
NFV	13992	15093	18429	28357	65080	103575

Tab. 1: $N = 50$.

	$N=50$				$N=200$			
Method	NIT	NFV	Fail	Time	NIT	NFV	Fail	Time
SBFGS	11256	12178	-	1.03	30429	36080	1	25.11
SBC2	11065	12670	-	1.05	32448	40019	1	26.34
SDFP	46010	48237	8	3.78	92799	100461	15	74.88
BFGS	14958	16474	1	1.26	36099	39991	2	27.21
BC2	12733	15152	1	1.04	30116	35814	2	23.88
DFP/1	79486	84215	35	6.66	146851	158979	32	113.75
DFP/2	15163	35422	2	1.84	36795	84255	2	42.15

Tab. 2:

results for standard VM methods: the BFGS method (BFGS), method from the Broyden class with $\eta = 2$ (BC2), both with scaling in the first iteration, the DFP method (DFP/1) and the DFP method with the third inequality in (1) replaced by $|g_{k+1}^T d_k| \le 0.1 |g_k^T d_k|$ (DFP/2), both without scaling.

This table demonstrates that the shifted BFGS method is more efficient than the standard BFGS method and the shifted DFP method can give better results than the standard DFP method with usual scaling strategies and usual line search methods (better scaling strategies for standard VM methods are introduced in [5]).

3. Limited-memory methods

These methods belong to shifted VM methods; they satisfy (2) with (positive semidefinite) matrix $A_k = U_k U_k^T$, where U_k, $k \ge 1$, is rectangular. We need to store only matrix U_k, which can be updated using relation $U_{k+1} = V_k U_k$, $k \ge 1$, with some matrix V_k. The shifted BFGS method (see Section 2.) is ideal as starting method, see [9]. Thus in this section we will assume that the starting iterations have been executed and that matrix U has $m \ge 1$ columns in all iterations.

With $V = I + pq^T$, good results were obtained only for $q = Bs$ and $q = y$. Thus we will investigate the case $V = I + p_1 y^T + p_2 s^T B,\ \ p_1 \in \mathcal{R}^N,\ \ p_2 \in \mathcal{R}^N$.

3.1. Methods based on general expression of the basic update

Denoting $\bar{\delta} = \bar{a} s^T BABs - \bar{b}^2$ ($\bar{\delta} \ge 0$ by the Schwarz inequality), we can write

$$A_+ = A + \varrho \tilde{s}\tilde{s}^T/\tilde{b} - Ayy^T A/\bar{a} + (q_2 q_2^T - v_2 v_2^T)/(\bar{a}\bar{\delta}), \qquad q_2^T y = 0,$$

for $\bar{\delta} \ne 0$ (which implies $\bar{a} \ne 0$), where $v_2 = \bar{a} ABs - \bar{b} Ay$ (see [9] for the case $\bar{\delta} = 0$). To construct this update, we choose vector parameter q_2 satisfying $q_2^T y = 0$ and calculate p_1 and p_2, using formulas $p_2 = (q_2 - v_2)/\bar{\delta},\ \ p_1 = \left(\sqrt{\varrho \bar{a}/\tilde{b}}\, \tilde{s} - Ay - \bar{b} p_2\right)\Big/\bar{a}$.

SSBC - simple method based on the shifted Broyden class

Surprisingly, very good results were obtained with choice $q_2 = \sqrt{\bar{\delta}}\left((\bar{a}/\tilde{b})\tilde{s} - Ay\right)$. Then we have the shifted BFGS update (3) with the additional term $-v_2 v_2^T/(\bar{a}\bar{\delta})$.

DSBC - method with direction vector after the shifted Broyden class

It suffices to compare value $H_+ Bs$, see [9].

3.2. Variationally-derived limited-memory methods

Standard VM methods can be obtained as updates with the smallest correction of VM matrix in the sense of some norm (see [5]). We extend this approach to limited-memory methods, replacing the shifted quasi-Newton condition $U_+ U_+^T y = A_+ y = \varrho\tilde{s}$ equivalently by (the first two conditions imply the third one)

$$U_+^T y = z, \qquad U_+ z = \varrho\tilde{s}, \qquad z^T z = \varrho\tilde{b}. \tag{5}$$

Theorem 3.1. *Let T be a symmetric positive definite matrix, $z \in \mathcal{R}^m$ and denote $\mathcal{U}$ the set of $N \times m$ matrices. Then the unique solution to (Frobenius matrix norm)*

$$\min\{\varphi(U_+) : U_+ \in \mathcal{U}\} \text{ s.t. } (5), \quad \varphi(U_+) = y^T T y \, \|T^{-1/2}(U_+ - U)\|_F^2,$$

is

$$U_+ = U - \frac{Ty}{y^T T y} y^T U + (\varrho\tilde{s} - Uz + \frac{y^T U z}{y^T T y} Ty)\frac{z^T}{z^T z}. \tag{6}$$

If $Ty = \varrho\tilde{s} - Uz$, then the value of $\varphi(U_+)$ reaches its minimum on the set of symmetric positive definite matrices T and update (6) can be written in the form

$$U_+ = U - [1/(\varrho\tilde{b} - y^T U z)](\varrho\tilde{s} - Uz)(U^T y - z)^T. \tag{7}$$

VAR1 - type 1 variationally-derived method

By analogy with the BFGS method, we set $z = \vartheta U^T B s$, $\vartheta = \pm\sqrt{\varrho\tilde{b}/\bar{c}}$ in (7):

$$U_+ = U - [1/(\varrho\tilde{b} - \vartheta\bar{b})](\varrho\tilde{s} - \vartheta ABs)\,(y - \vartheta Bs)^T U,$$

which gives the best results for the choice $\mathrm{sgn}(\vartheta\bar{b}) = -1$ (compare with Theorem 3.2).

VAR2 - type 2 variationally-derived method

With ϑ, z given as above and with the simple choice $Ty = \tilde{s}$, (6) leads to

$$U_+ = U - \tilde{s} y^T U/\tilde{b} + \left[\left(\varrho/\vartheta + \bar{b}/\tilde{b}\right)\tilde{s} - ABs\right] s^T B U/\bar{c}.$$

Efficiency of both these methods significantly depends on the value of parameter ϱ. Very good results were obtained with the following choices of ϱ: $\varrho^{(1)} = \nu$, $\varrho^{(2)} = \sqrt{\nu\varepsilon}$, $\varrho^{(3)} = \zeta/(\zeta + \zeta_+)$ and $\varrho^{(4)} = \sqrt{\mu\sqrt{\varrho^{(3)}/2}}$, where $\nu = \mu/(1-\mu)$ and $\varepsilon = \sqrt{\zeta\hat{a}/a}$.

3.3. Global convergence

Theorem 3.2. *Consider methods SSBC, DSBC, VAR1 and VAR2 satisfying Assumption 2.2 with μ sufficiently small. Let the line search method fulfil (1) and the objective function satisfy Assumption 2.1. If $\mu^2 \le \zeta \hat{a}/a$ or $\mu = \underline{\mu}$ and, for VAR1, $\vartheta_k = -\mathrm{sgn}\bar{b}_k \min\left[\tilde{C}, \sqrt{\varrho_k \tilde{b}_k/\bar{c}_k}\,\right]$, $k \ge 1$, for some $0 < \tilde{C} < \infty$, then* $\liminf_{k\to\infty} |g_k| = 0$.

3.4. Computational experiments

We use the collections of problems [7] (Test 28) and [6] (Test 14, usually well-conditioned problems), the final precision $\|g(x^\star)\|_\infty \le 10^{-6}$, $m = 10$, the choice of μ after (4) with corrections and the shifted BFGS method for starting iterates.

Symbols NFV and 'Fail' have the same meaning as in Section 2.2. The first four rows of the table give results for methods SSBC, DSBC, VAR1 and VAR2. For methods VAR1, VAR2 we use $\varrho = \varrho^{(4)}$ for VAR1 in Test 28 and $\varrho = \varrho^{(3)}$ otherwise. The last four rows contain results for the following limited-memory methods: NS (see [8]), BNS (see [1]), RH (see [4]) and CGM (see [3]); this method often stopped before the requested precision was achieved. Note that methods BNS and NS store $2m$ vectors while method CGM stores no additional vectors.

	Test 28, 80 problems			Test 14, 22 problems					
	$N = 1000$						$N = 5000$		
Method	NFV	Fail	Time	NFV	Fail	Time	NFV	Fail	Time
SSBC	104246	-	1:06.8	20203	-	15.74	83866	-	13:21.4
DSBC	109178	-	1:13.3	21969	-	17.25	95090	-	14:48.0
VAR1	99261	-	1:00.3	19680	-	13.86	72214	-	9:32.8
VAR2	92699	-	0:55.4	18546	-	13.76	70127	-	9:30.6
NS	98275	-	0:51.9	21456	-	15.17	84426	-	11:02.2
BNS	122593	1	1:02.8	26003	1	16.55	77803	-	11:38.1
RH	113925	-	0:56.1	33181	-	24.09	150827	2	12:34.6
CGM	223219	2	1:15.3	41049	-	17.91	168471	1	6:45.3

Tab. 3:

References

[1] R.H. Byrd, J. Nocedal, R.B. Schnabel: *Representation of quasi-Newton matrices and their use in limited memory methods*. Math. Programming **63**, 1994, 129–156.

[2] R. Fletcher: *Practical methods of optimization.* John Wiley & Sons, Chichester, 1987.

[3] J.Ch. Gilbert, J. Nocedal: *Global convergence properties of conjugate gradient methods for optimization.* SIAM J. Optim. **2**, 1992, 21–42.

[4] P.E. Gill, M.W. Leonard: *Limited-memory reduced-Hessian methods for large-scale unconstrained optimization*. SIAM J. Optim. **14**, 2003, 380–401.

[5] L. Lukšan, E. Spedicato: *Variable metric methods for unconstrained optimization and nonlinear least squares*. J. Comput. Appl. Math. **124**, 2000, 61–95.

[6] L. Lukšan, J. Vlček: *Sparse and partially separable test problems for unconstrained and equality constrained optimization*. Report V-767, Prague, ICS AS CR, 1998.

[7] L. Lukšan, J. Vlček: *Test problems for unconstrained optimization*. Report V-897, Prague, ICS AS CR, 2003 (`www.cs.cas.cz/~luksan/test.html`).

[8] J. Nocedal: *Updating quasi-Newton matrices with limited storage*. Math. Comp. **35**, 1980, 773–782.

[9] J. Vlček, L. Lukšan: *New variable metric methods for unconstrained minimization covering the large-scale case*. Report V-876, Prague, ICS AS CR, 2002.

[10] J. Vlček, L. Lukšan: *Additional properties of shifted variable metric methods*. Report V-899, Prague, ICS AS CR, 2004 (`www.cs.cas.cz/~luksan/reports.html`).